Pergamon

0042-6989(94)00113-8

Vision Res. Vol. 35, No. 2, pp. 303–312, 1995

0042-6989/95 $7.00 + 0.00

Interobserver Agreement for Measurement of Grating Acuity and Interocular Acuity Differences with the Teller Acuity Card Procedure*

CLAY MASH,†‡ VELMA DOBSON,†§ NANCY CARPENTER§

Received 12 July 1993; in revised form 16 March 1994

Interobserver reliability of the Teller acuity card (TAC) procedure for estimating acuity and interocular acuity differences (IADs) was assessed with 342 infants and children who had been treated in a neonatal intensive care unit for preterm birth and/or perinatal complications. Subjects were tested binocularly at term and monocularly at 4, 8, 11, 17, 24, 30, 36, and 48 months corrected age with TACs. Testers were masked to the location and spatial frequency of the grating on each card. Of the interobserver test–retest scores, 67% differed by no more than 0.5 octave, and 87% of the test pairs differed by no more than 1 octave. Of the test–retest comparisons of a subject's IAD, 54% showed agreement of 0.5 octave or better, and 76% differed by no more than 1 octave. Interobserver agreement for binocular and monocular tests was similar to that reported previously for visually and neurologically at-risk infants and children tested with the forced-choice preferential-looking procedure or with prototype acuity cards. Interobserver agreement for IAD estimates was somewhat less than that reported for a sample of infants with ocular disorders. There were no systematic differences in interobserver agreement between eyes tested first and eyes tested second, nor was interobserver agreement related to subject's medical diagnosis. Interobserver agreement was influenced, however, by the spatial frequencies of the particular gratings used during testing and, to a limited extent, by the age of the child. The duration of individual tests and observers' ratings of confidence in their acuity estimate were not reliable indicators of test–retest pairs that were not in agreement. The results demonstrate the reliability of the TAC procedure, but suggest that acuity estimates critical to a patient's diagnosis or treatment should be confirmed by repeat testing.

Visual acuity Grating acuity Interocular acuity differences Interobserver reliability Acuity cards Infants Children Perinatal complications

INTRODUCTION

The acuity card procedure (McDonald, Dobson, Sebris, Baitch, Varner & Teller, 1985; Teller, McDonald, Preston, Sebris & Dobson, 1986) is a rapid, subjective method for estimating grating acuity in infants and young children in clinical settings. In the procedure, the child is shown a series of gray cards, each containing a black-and-white square-wave grating located to the left or right of a small central aperture. The tester, who is unaware of the location of the grating on each card, watches the child's eye movements through the aperture and decides, on the basis of the child's looking behavior, which cards contain gratings that can be resolved by the child. Acuity is estimated as the spatial frequency of the finest grating that the tester judges that the child can resolve, as indicated by the child's consistent looking toward the location of the grating upon repeated presentations of the card.

Because the acuity card procedure depends on the subjective judgment of the tester, it is important that studies of the procedure's reliability and validity be conducted. Initial studies, conducted using a prototype version of the acuity cards, indicated that both intra-observer test–retest reliability (McDonald *et al.*, 1985; Hertz, 1987, 1988; Hertz & Rosenberg, 1988, 1992) and interobserver test–retest reliability (McDonald *et al.*, 1985; McDonald, Ankrum, Preston, Sebris & Dobson, 1986a; McDonald, Sebris, Mohn, Teller & Dobson, 1986b; Preston, McDonald, Sebris, Dobson & Teller,

*Presented in part at Optical Society of America, Topical Meeting on Noninvasive Assessment of the Visual System, Santa Fe, N.M., 4–7 February 1991.

†Department of Psychology, Langley Hall, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.

‡To whom all correspondence should be addressed.

§Department of Psychiatry, Langley Hall, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.

1987; Hertz & Rosenberg, 1988, 1992; Hertz, Rosenberg, Sjo & Warburg, 1988; Dobson, Carpenter, Bonvalot & Bossler, 1990; Heersema & van Hof-van Duin, 1990) of the acuity card procedure are similar to that of the scientifically rigorous forced-choice preferential-looking (FPL) procedure used for laboratory testing of acuity in infants and young children. Commercial production of the acuity card [Teller Acuity Cards (TAC), Vistech, Inc., Dayton, Ohio] led to widespread use of the cards in clinical settings. However, it is unclear whether studies of reliability conducted with prototype cards are applicable to the Teller cards, because of differences in stimulus configuration between the two card types (Robinson, Moseley & Fielder, 1988; Hainline, Evelyn & Abramov, 1989; cf. Dobson & Luna, 1993). Initial results from a group of children with perinatal complications, tested between 1 and 24 months of age with the Teller cards, showed interobserver reliability similar to that of a similar group of children tested with prototype cards (Dobson & Carpenter, 1991).

The present study continued the investigation of interobserver reliability of Teller card assessment of grating acuity in infants and young children by using a larger sample and an extended age range that included children between one month and four years of age. Additionally, the interobserver reliability of estimates of each child's interocular acuity difference (IAD) was examined. To our knowledge, this is the first study to report interobserver reliability across the age span between birth and 4 yr, and only the second to examine the agreement between testers in the estimation of IADs for children in this age range.

METHOD

Subjects

Subjects were 342 children treated in the Neonatal Intensive Care Unit (NICU) of Magee-Womens Hospital, Pittsburgh, Pa, and who were born between December 1984 and September 1991. All were participants in a longitudinal study of visual acuity and visual field development of infants treated in the NICU. Seventy-five children had no complications other than preterm birth. These healthy preterm "control" subjects had a gestational age $\leqslant 36$ weeks, without any of the following conditions: mechanical ventilation >24 hr, supplemental oxygen $\geqslant 40\%$, exchange transfusion, sepsis, retinopathy of prematurity, central nervous system malformations, progressive hydrocephalus, meningitis, cyanotic cardiac malformations, or multiple congenital anomalies.

The remaining 267 children had one or more of the following perinatal complications: bronchopulmonary dysplasia (BPD), hyaline membrane disease (HMD), neonatal hypoxia/asphyxia, periventricular leukomalacia (PVL), intraventricular hemorrhage (IVH), retinopathy of prematurity (ROP), or persistent pulmonary hypertension or the newborn accompanying full-term birth.

Children were tested longitudinally at one or more of the following ages: (a) term [−3 through 31 days corrected age (age from due date)]; (b) 4 months (80–143 days); (c) 8 months (223–283 days); (d) 11 months (306–382 days); (e) 17 months (474–535 days); (f) 24 months (699–761 days); (g) 30 months (887–944 days); (h) 36 months (1069–1125 days); and (i) 48 months (1432–1490 days). Only four subjects were tested at all nine ages. The numbers of subjects tested at eight, seven, six, five, four, three, two, and one age(s) were, respectively: 6, 12, 22, 43, 54, 55, 66, and 80.

Apparatus

The apparatus consisted of a set of 19 Teller acuity cards and an acuity card screen (Vistech, Inc., Dayton, Ohio), which contained a 20.2×46.8 cm aperture through which the cards were presented. Fifteen of the cards contained gratings ranging in spatial frequency from 0.32 to 38 c/cm, in approximately half-octave steps. The 16th card was a blank, gray card. The remaining three cards each contained a 0.32 c/cm grating. Inclusion of the extra three 0.32 c/cm cards provided enough cards to allow 10 different subsets of 10 cards, each spanning a different range of highest and lowest spatial frequencies (see Table 1). Random selection of the subset to be used for each test ensured that testers were masked to the absolute spatial frequencies of the gratings that were shown to the child during testing.

TABLE 1. Grating subsets*

Grating spatial frequency (c/cm)	1	2	3	4	5	6	7	8	9	10
0.32	x									
0.32	x	x								
0.32	x	x	x							
0.32	x	x	x	x						
0.43	x	x	x	x	x					
0.64	x	x	x	x	x	x				
0.86	x	x	x	x	x	x	x			
0.30	x	x	x	x	x	x	x	x		
1.60	x	x	x	x	x	x	x	x	x	
2.40	x	x	x	x	x	x	x	x	x	x
3.20		x	x	x	x	x	x	x	x	x
4.80			x	x	x	x	x	x	x	x
6.50				x	x	x	x	x	x	x
9.80					x	x	x	x	x	x
13.0						x	x	x	x	x
19.0							x	x	x	x
26.0								x	x	x
38.0									x	x
Blank										x

*Cards comprising the 10 possible subsets used during testing. Infants at term and 4 months were tested with subsets 1–7, at a distance of 31 cm, so that grating spatial frequencies ranged from 0.19 through 11 c/deg. At 8 and 11 months, infants were tested at 31 cm with subsets 4–10, with spatial frequencies ranging from 0.19 to 22 c/deg. From 17 through 36 months, children were tested at 55 cm with subsets 4–10, with spatial frequencies between 0.31 and 38 c/deg. At 48 months, subsets 4–10 were used at a distance of 84 cm, with spatial frequencies ranging from 0.47 to 57 c/deg.

Procedure

After informed consent had been obtained from the parents, the child was held or seated in front of the acuity card screen. Test distance was 31 ± 3 cm for infants between birth and 12 months of age, 55 ± 3 cm for children aged 17 and 36 months, and 84 ± 3 cm for children at 48 months.

At each test age, each child's acuity was measured by two independent observers, using the procedure described previously for a study of interobserver reliability of prototype cards (Dobson *et al.*, 1990). At the initial test age (term), each observer conducted a test of binocular acuity, because most infants would not remain awake for monocular testing. At all other ages, monocular acuity was tested according to the following order:

Eye tested first:
(1) observer 1 tests one eye (e.g. left);
(2) observer 2 tests the same eye.
Eye tested second:
(3) observer 2 tests remaining eye (e.g. right);
(4) observer 1 tests the same eye.

Testers were aware that the spatial frequencies of the gratings on the cards progressed from lower to higher frequencies, but were masked, through the use of randomly-selected subsets of cards, to the absolute spatial frequencies of the gratings on each card. Testers were also masked to the location of the grating on each card and to the results obtained by the other tester. A low spatial frequency card and a blank card were available to the tester at all times to allow the tester to observe the subject's response when a grating was clearly present, and when the grating was absent.

Each tester's estimate of the finest grating that the child could resolve, each tester's confidence in the accuracy of that acuity estimate (on a 5-point scale), and the time required for each test were recorded.

Data analysis

Analyses of monocular test results were conducted separately for data from the eye tested first and data from the eye tested second, in order to avoid problems of lack of independence that arise when statistical analyses are conducted on data of two eyes from a single object.

Interobserver agreement for a binocular or a monocular test was calculated as the difference in octave units between acuity estimates obtained by the two testers (one octave corresponds to a halving or doubling of grating spatial frequency). For evaluation of interobserver agreement concerning a child's IAD, each tester's estimate of the child's IAD was calculated as the algebraic difference in octaves between the acuity value obtained for the child's right eye and the acuity value obtained for the child's left eye. The difference in octaves between the IADs obtained by the two testers provided a measure of interobserver agreement for the child's IAD. For example, if observer 1 found that the acuity of the right eye was 1.0 octave better than the acuity of the left eye, and observer 2 found that the acuity of the right eye was 0.5 octave worse than the acuity of the left eye, then the two observers differed by a total of 1.5 octaves in their estimates of the child's IAD.

For analyses involving actual acuity scores (e.g. analyses of observer differences in mean acuity estimates), log acuity scores were used.

Although subjects were tested longitudinally, few were tested at all nine test ages. For most analyses, we therefore conducted separate tests for each age and used the Bonferroni correction to determine the α level. Because in some cases this strategy resulted in relatively small α levels (down to 0.003), we report contrasts (and associated α levels) that closely approach significance, along with contrasts that are significant. Finally, because there are no prescribed standards for interobserver reliability, we report results for the two levels of interobserver agreement most frequently presented in previously published reports: test pairs differing by ≤ 1.0 octave, and the more stringent criterion of differences no greater than 0.5 octave.

RESULTS

Interobserver agreement for binocular and monocular tests

Figure 1 presents the percentage of test–retest pairs on which observers' estimates of acuity differed by 0, 0.5, 1.0, or >1.0 octave at each test age. For binocular tests at term, 66% of test–retest pairs differed by no more than 0.5 octave, and 87% differed by no more than 1.0 octave. Across all pairs of monocular acuity estimates, 67% of test–retest pairs differed by 0.5 octave or less, and 87% differed by no more than 1.0 octave. Mean test–retest difference was 0.6 octave (SD = 0.6) across all ages, and ranged from 0.4 octave (SD = 0.3) at 48 months (for the eye tested second) to 0.7 octave (SD = 0.6) at 4 months (for the eye tested second). The median test–retest difference was 0.5 octave across all ages, and also 0.5 octave at each test age, for eyes tested first and eyes tested second.

Discrepant test–retest pairs: possible sources and indicators of disagreement

The finding that two-thirds (67%) of test–retest pairs of binocular and monocular tests differed by no more than 0.5 octave indicates that TAC observers agree within a very narrow range on most estimates. Furthermore, the finding that 87% of test–retest pairs differed by no more than 1.0 octave indicates that it is rare for observers to differ by more than two acuity cards in their estimates of acuity. However, because clinical testing focuses on individual patients, it is important to determine whether there are identifiable subject or test characteristics that would help to identify test results that are likely to be unreliable.

Subject characteristics. The hypothesis investigated in this analysis was that acuity results would be more reliable for subjects who were easy to test than for subjects who were difficult to test. Ease of testing might

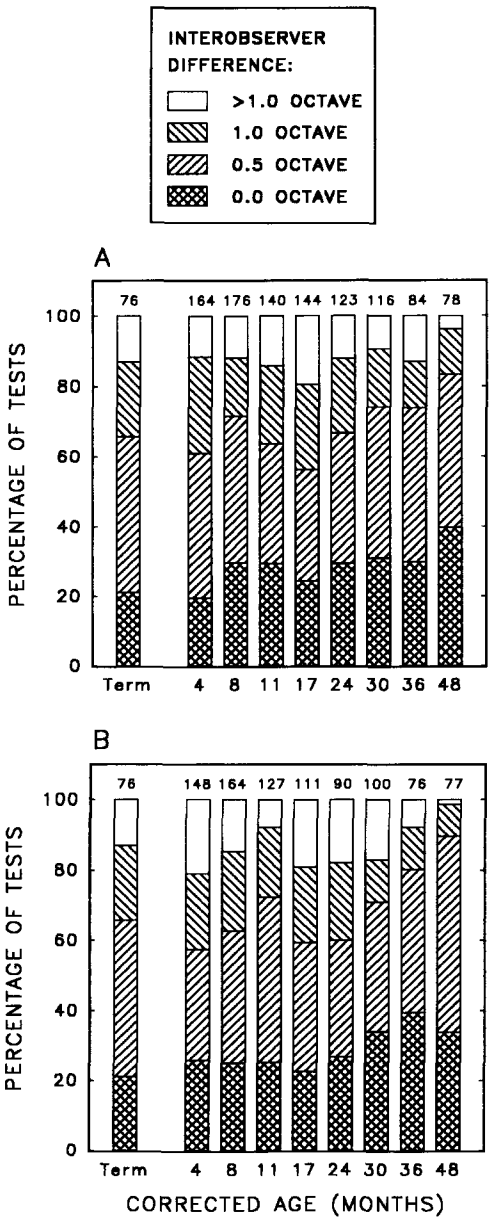

FIGURE 1. Percentage of interobserver test–retest pairs with acuity scores that differed by 0, 0.5, 1.0 and >1.0 octave, for binocular tests at term and for (A) the eye tested first and (B) the eye tested second at each of the eight older test ages. Numbers above each bar indicate the number of test–retest pairs represented by the bar.

relate to (i) presence or absence of perinatal complications affecting visual or neurological development; (ii) age at testing; or (iii) whether the eye was tested first or second (i.e. the influence of fatigue).

To evaluate the possibility that interobserver agreement is poorer in children with severe perinatal complications than in children with minimal complications, results from two groups of children were compared. The first (high risk) group consisted of preterm children with PVL or stage 3 or 4 ROP, and fullterm, perinatally asphyxiated children. The second group consisted of healthy preterm "control" children. Two series of χ^2 analyses compared the frequency of "good" vs "poor" interobserver agreement between the two groups at each age. For one series of analyses, "good" agreement was defined as a difference $\leqslant 0.5$ octave. For the other series, "good" agreement included differences $\leqslant 1.0$ octave. Test results, evaluated according to family-wise error rate, yielded no significant difference—at either level of agreement—between children in the high risk group and children in the healthy preterm group for the proportion of test–retest pairs with good vs poor interobserver agreement.

Another possible covariate of testing difficulty is the age of the child, since compliance with monocular testing varies as a function of age (Sebris, Dobson, McDonald & Teller, 1987). Examination of Fig. 1 shows that agreement appears poorest for eyes tested second at the 4-month age and best for eyes tested second at the 48-month age. To examine whether interobserver agreement differs reliably as a function of the child's age, a total of six repeated-measures ANOVAs were conducted, with overall Fs evaluated according to family-wise error rates. Because very few children were assessed at each of the eight monocular test ages, analyses including three different subsets of test ages were carried out, and were done so independently for eyes tested first, and eyes tested second. For each eye (i.e. those tested first, and those tested second), two analyses included four consecutive test ages (4, 8, 11, and 17 months; and 24, 30, 36, and 48 months). A third ANOVA was carried out for each eye that included four test ages that were distributed across a wider range: 8, 17, 30, and 48 months. Each pair of analyses involved a different subgroup of the total subject sample, although none were mutually exclusive.

Among test pairs for eyes tested first, interobserver agreement did not differ across any of the test age distributions. For eyes tested second, there were no significant differences in mean interobserver difference across the age range between 4 and 17 months, but mean difference did approach significance when compared across the 24- to 48-month range ($F_{3,69} = 4.2$, $P < 0.01$, $\alpha = 0.008$). Paired comparisons revealed that mean interobserver difference approached significance for the comparison between 24 and 48 months, with agreement being better at the 48-month test age ($t_{23} = 3.0$, $P < 0.01$, $\alpha = 0.008$). Interobserver agreement for eyes tested second did not differ significantly across the wider distribution of test ages.

A third factor that may affect ease of testing is whether an eye was tested at the beginning or at the end of the test session. If fatigue makes testing more difficult, interobserver agreement would be predicted to be poorer for the eye tested second than for the eye tested first. χ^2 analyses, however, yielded no significant difference, at either the half- or full-octave level of agreement, in frequency of good agreement for test–retest data from the eye tested first vs the eye tested second.

Test characteristics. These analyses were conducted to evaluate whether test characteristics likely to correlate with difficulty of testing could serve as indicators of test reliability. The two test characteristics analyzed were test duration and observer's confidence in the acuity results.

Children who are difficult to test often require longer test times than children who are easy to test. Therefore, it was hypothesized that longer test duration would be associated with poor interobserver agreement. To test this hypothesis, the longer time for each pair of tests was selected, and t-tests were conducted to compare test duration of the lengthier test in test–retest pairs with good agreement versus test duration of the lengthier test in pairs with poor agreement at each age. The analyses revealed that for eyes tested first at 24 months, test pairs with a difference >0.5 octave had longer durations—approaching significance—relative to those with differences no greater than 0.5 octave ($t_{121} = 2.9$, $P < 0.005$, $\alpha = 0.003$). At no other ages was test duration related to interobserver agreement. Also, when good agreement was defined to include test pairs differing by a full octave or less, there were no significant differences in test duration between levels of agreement at any age.

Observers' ratings of confidence in their acuity estimates may be another index of testing difficulty. The lower confidence rating from each test pair was selected, and χ^2 tests were conducted to compare the frequencies of low (1, 2, or 3 on a 5-point scale) and high (4 or 5) confidence ratings on test pairs with good vs poor interobserver agreement. χ^2 analyses revealed only one instance of significant association between observers' confidence and interobserver agreement. For eyes tested first at 8 months, test pairs characterized by at least one rating of low confidence were more likely to also reflect poor interobserver agreement ($\chi^2_1 = 10.5$, $P < 0.002$, $\alpha = 0.003$), when good agreement was defined to include differences ≤ 0.5 octave. When good agreement was defined to include differences ≤ 1 octave, there were no significant associations between observer confidence and interobserver agreement.

Bias related to card subsets. To keep testers masked to the absolute spatial frequencies of the gratings used in testing, the starting spatial frequency was varied across tests through the use of different subsets of cards (Table 1). Analyses showed that across all tests and across all ages, acuity values were significantly correlated with the subsets of cards used ($r = 0.50$, $P < 0.001$); the lower the spatial frequencies in the subset of cards, the lower was the acuity estimate. The largest effect of card subset was found at 8 months [Fig. 2(A)]. At this age, an average difference of just over 1 octave was found between acuity values obtained with the subset containing the lowest vs the subset containing the highest spatial frequencies. These extreme subsets differed by 3 octaves in the spatial frequency of the initial card in the subset, as shown in Table 1. The smallest effect of card subset was found at 48 months [Fig. 2(B)], where the difference between acuity values obtained using the subset containing the lowest vs the highest spatial frequencies averaged about 0.5 octave.

Divergence in spatial frequency of the starting card between observers may therefore have contributed to disagreement within observer pairs. To examine the effect of card subset bias on interobserver agreement, the difference in octaves between the spatial frequencies of the starting cards for each pair of tests was calculated. t-Tests were conducted to determine whether the difference in spatial frequency between starting cards was greater for test–retest pairs that differed by more than 0.5 octave than for those that differed by 0.5 octave or less. The results showed that test pairs that differed by more than 0.5 octave had significantly greater start-card differences than those characterized by agreement of 0.5 octave or better among eyes tested first at 8 and 11 months ($t_{174} = -3.6$, $P < 0.001$; and $t_{138} = -3.6$, $P < 0.001$, $\alpha = 0.003$, respectively) and eyes tested second at 8 months ($t_{162} = -3.1$, $P < 0.003$, $\alpha = 0.003$). Differences between levels of agreement as a function of card subset approached significance among eyes tested second at 36 months ($t_{74} = -2.9$, $P < 0.006$, $\alpha = 0.003$). Using the 1-octave criterion for good agreement, the start-card difference approached significance only among eyes tested second at 8 months ($t_{162} = -3.1$, $P < 0.01$, $\alpha = 0.003$).

Observer bias. Another factor that could contribute to poor interobserver agreement is observer bias, or the tendency for different testers to use slightly different criteria in estimating acuity. Differences among testers have been reported for FPL testing of normal infants

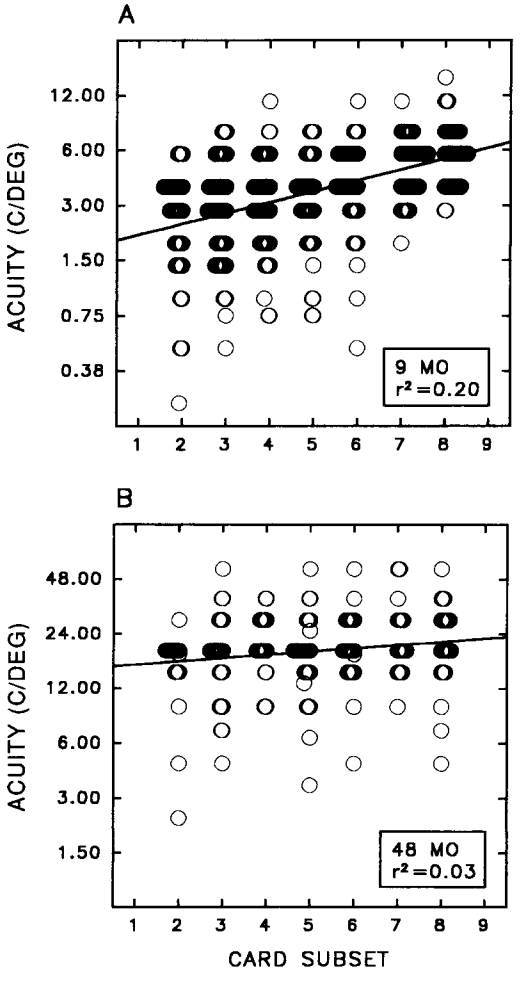

FIGURE 2. Acuity estimates, both for eyes tested first and eyes tested second, plotted as a function of the acuity card subset that was used for each test of subjects at (A) 8 months, and (B) 48 months.

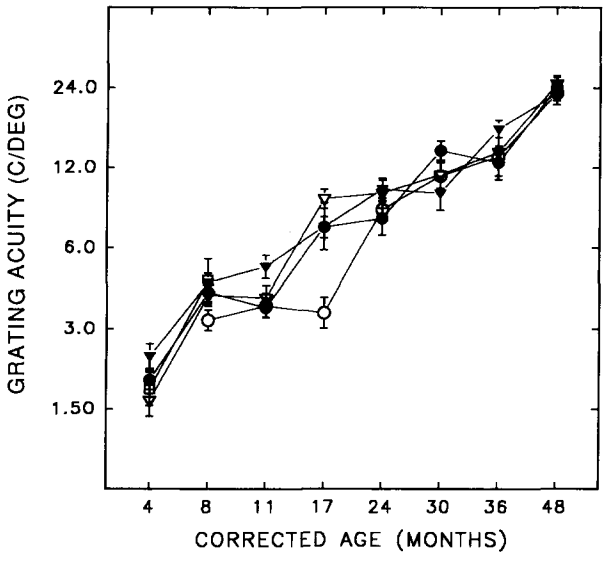

FIGURE 3. Mean acuity scores (first- and second-eye tests combined) obtained by the study's five primary testers from children in the healthy preterm group. Data are presented only for those ages at which a given tester completed at least eight acuity estimates. Symbols represent testers 1 (○), 2 (●), 3 (▽), 4 (▼), and 5 (□). Differences among testers were significant only at 17 months, where the mean score for observer 1's first-eye tests was lower than the mean score of each of the other testers.

(Teller, Mar & Preston, 1992) and for acuity card testing of patients in a clinical setting (Quinn, Berlin & James, 1993). Mohn, van Hof-van Duin, Fetter, de Groot and Hage (1988), however, reported a lack of any significant difference among the mean acuity estimates of two to four acuity card testers, for five test ages.

To see if acuity results differed significantly among testers, we calculated each tester's mean acuity score for each test age. Only acuity results from subjects in the healthy preterm group were used, to eliminate bias that would arise if one tester had tested a high percentage of subjects with below-normal acuity. In addition, mean acuity scores for a tester were included in the analysis only for the test ages at which the tester had completed at least eight acuity tests. The results of this comparison (Fig. 3) showed considerable similarity among acuity results of the five primary testers in the study. A one-way ANOVA was conducted for each eye at each test age, with overall F ratios evaluated according to family-wise error rate. Significant differences in mean acuity scores were found only for the eye tested first at 17 months ($F_{3,42} = 10.6$, $P < 0.0001$, $\alpha = 0.004$). Differences in mean acuity scores approached significance for the eye tested second at 17 months ($F_{2,23} = 6.3$, $P < 0.007$, $\alpha = 0.004$). *Post hoc* contrasts revealed that observer 1's mean score was significantly lower than the mean score of each of the other observers at 17 months (see Fig. 3).

If observer bias has a significant effect on inter-observer agreement, one would expect to find that certain pairs of testers (e.g. a tester who tends to report high acuity values and a tester who tends to report low acuity values) are more likely than other pairs of testers to show large differences in acuity scores. Figure 4 provides agreement data for the pair of observers whose mean test–retest difference was the highest [Fig. 4(A)], and for the pair whose mean test–retest difference was the lowest [Fig. 4(B)].

While examination of Fig. 4 seems to suggest otherwise, an analysis was conducted to determine whether some observer pairs showed an unusually high frequency of poor interobserver agreement. For this analysis, data of all subjects were used, and the frequency of good vs poor agreement for different pairings of testers was examined. χ^2 analyses were conducted to compare particular observer pairs vs all other pairs, for each of the six possible pairings of the four testers with the highest numbers of tests. None of the χ^2 values were significant at either level of agreement, suggesting that observer

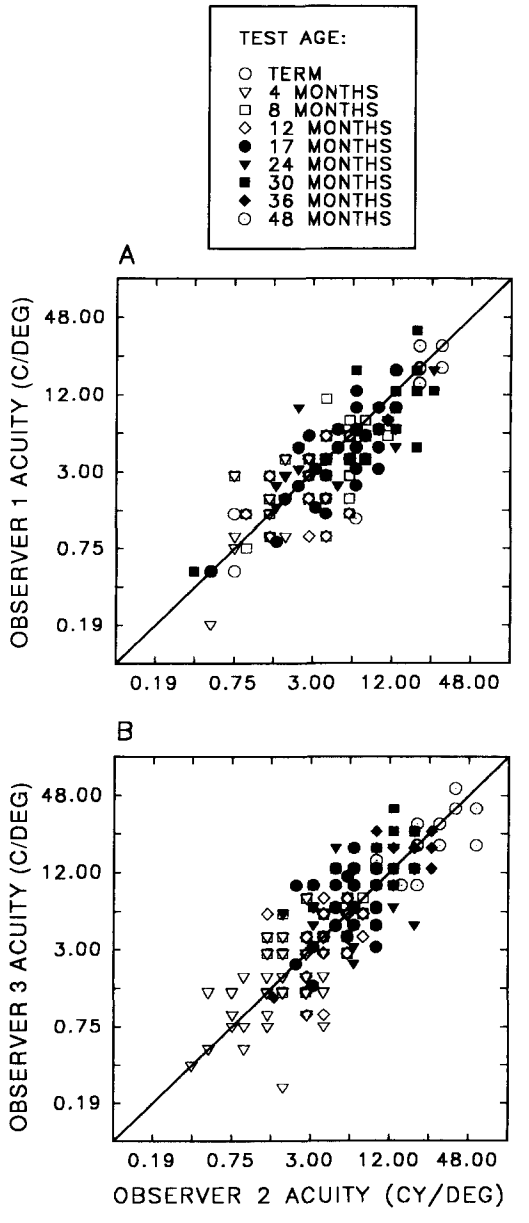

FIGURE 4. Test–retest acuity estimate pairs for examinees tested by observers 1 and 2, whose mean interobserver difference was the *least* of all possible pairings of the four most frequent observers (A), and those for examinees tested by observers 2 and 3, whose mean difference was the *greatest* (B).

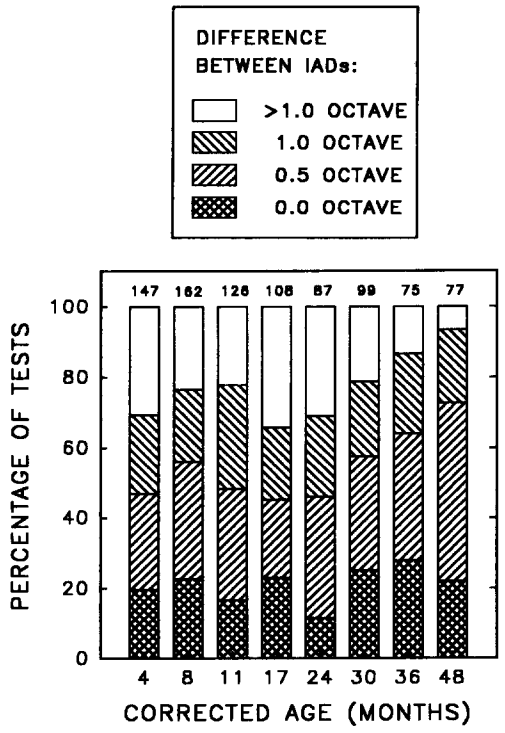

FIGURE 5. Interobserver agreement for estimates of within-subject interocular acuity differences (IADs) for the eight ages at which monocular acuity was assessed. Numbers above each bar indicate the number of IAD comparisons represented by each bar. The percentage of comparisons that differed by no more than 1 octave ranged from 66% at 17 months to 94% at 48 months.

pairing has, at most, a very limited effect on interobserver agreement.

Underestimation vs overestimation of acuity. A final analysis attempted to determine whether overestimation of acuity or underestimation of acuity was more likely to occur in instances of poor interobserver agreement. For this analysis, the median acuity value at each test age for the healthy preterm group was determined. Then, using only data from the healthy preterm group, each acuity score in the 88 test pairs in which observers' scores differed by more than 1.0 octave was compared with the median acuity value for the group at that age. In 62 of the test pairs, both acuity scores were within 1 octave of the median score for that age. One acuity score was more than 1 octave below (poorer than) the age-appropriate median in 19 (73%) of the pairs, while one acuity score was more than 1 octave above (better than) the age-appropriate median in seven (27%) of the pairs. Thus, while both under- and overestimation of acuity occurred, underestimation was found more frequently.

Interobserver agreement for estimation of IADs

Figure 5 shows interobserver agreement for estimation of IADs, for all ages at which monocular acuity was tested. Across all ages, 54% of IAD test–retest comparisons differed by 0.5 octave or less, and 76% differed by no more than one octave.

Examination of Fig. 5 reveals that IAD test–retest agreement was lowest at 17 months and highest at 48 months. Three repeated-measures ANOVAs were conducted to examine whether IAD agreement differs reliably as a function of age. These analyses employed the same subsets of subject age that were used in the preceding age analyses as the within-subjects measures, and the absolute difference between the IAD estimates obtained by observer pairs for each subject who provided monocular data as the dependent measure. None of the analyses revealed significant age differences across the ranges examined.

χ^2 analyses were conducted to examine effects of the presence of perinatal complications on interobserver agreement on IAD estimates. No statistically significant relation between the presence of perinatal complications and interobserver difference in IAD was observed at either level of agreement.

DISCUSSION

The results of the present study indicated interobserver agreement of 0.5 octave or better in 67% of TAC test–retest comparisons, and agreement of 1 octave or better in 87% of test–retest comparisons of 1- to 48-month-old children who were treated in an NICU for preterm birth or perinatal complications. Also, interobserver agreement within 0.5 octave was obtained in 54% of observer-paired estimates of IAD for 4- to 48-month-old children, while agreement of 1 octave or better was found in 76% of between-observer IAD comparisons.

Comparison with previous results

Table 2 compares the results of the present study with the results of previous studies that examined interobserver test–retest reliability in subjects tested with the acuity card or the FPL procedure. Studies of normal infants and young children have shown agreement of 0.5 octave or better in 86–92% of test–retest comparisons, and agreement of 1 octave or better in 86–100% of test–retest comparisons (Atkinson, Braddick & Pimm-Smith, 1982; McDonald *et al.*, 1985, 1986a, b; Maurer, Lewis & Brent, 1989; Heersema & van Hof-van Duin, 1990).* Somewhat lower interobserver agreement was found in studies of infants and children with or at-risk for ocular or neurological abnormalities (Preston *et al.*, 1987; Hertz, 1988; Hertz & Rosenberg, 1988; Hertz *et al.*, 1988; Maurer *et al.*, 1989; Dobson *et al.*, 1990; Dobson & Carpenter, 1991). In at-risk or impaired subjects, the percentage of test–retest comparisons showing interobserver agreement of 0.5 octave or better ranged from 25% to 95%, and the percentage showing agreement of one octave or better ranged from 75% to 100%.

The interobserver agreement values of the present study are within the range reported previously for children with or at-risk for abnormalities. The present

*Heersema and van Hof-van Duin (1990) used a 0.3 octave step size between cards. This could potentially yield more precise estimates of acuity than those obtained with the between-card step size of 0.5 octave used in most studies. These investigators' values for interobserver agreement, however, are not substantially different from those of other studies, as shown in Table 2.

results are similar to those reported previously for a similar population tested with prototype acuity cards (Dobson *et al.*, 1990) and also to results reported by Hertz *et al.* for children with neurological abnormalities tested with the acuity card procedure (Hertz, 1988; Hertz & Rosenberg, 1988). Our results show lower percentages of agreement than those reported by Preston *et al.* (1987) for young infants with ocular abnormalities, but higher percentages than those reported by Maurer *et al.* (1989) for a sample of aphakic children, and higher percentages than those reported by Hertz *et al.* (1988) for a sample of mentally retarded, cortically visually impaired children.

The present study is only the second to provide data on interobserver reliability of estimating IADs in infants and young children. Both studies showed lower rates of agreement for estimating IADs than for estimating monocular or binocular acuity. This is not surprising, because the variability associated with each IAD comparison pair arises from four acuity tests (one test by each observer for the right eye and one test by each observer for the left eye), whereas the variability associated with each binocular or monocular test–retest comparison arises from only two acuity tests. Comparison of the results of the present study with those of the previous study (Preston *et al.*, 1987) shows lower percentages of agreement in the present study. This may be related to the greater variation in test ages and in medical diagnoses in the present study.

Factors associated with discrepant interobserver comparisons

The finding that interobserver agreement is 0.5 octave or better in 67% of test–retest comparisons (with 87% of test pairs differing by no more than 1 octave) and that 54% of IAD estimate pairs showed agreement of 0.5 octave or better (with 76% of IAD estimates differing by no more than 1 octave) supports the validity of the TAC as a measure of acuity in infants and young children. However, because moderate proportions of acuity and IAD estimates did not agree, caution should be observed when interpreting the results of any single estimate of acuity or IAD. The aim of many of the analyses in the present report was to determine whether there were characteristics of individual examinees or characteristics of individual acuity tests that might serve as indicators of inaccurate acuity tests.

Examinee characteristics evaluated were (i) presence of severe perinatal complications, (ii) whether an eye was treated first or second during the test session, and (iii) age at the time of the test. Although testers anecdotally report that children with ocular or neurological abnormalities are often difficult to test, our results showed that

TABLE 2. Interobserver agreement in acuity card and FPL studies

Study	Procedure	Step size (octaves)	Condition	Age	n*	≤0.5 octave	≤1.0 octave
Mash *et al.* NICU-treated	Acuity card	0.5	Binocular	1 month	78	66%	87%
McDonald *et al.* (1985) normals	Acuity card	1.0	Binocular	1 month	15	—	87%
Dobson *et al.* (1990) NICU-treated	Acuity card	0.5	Binocular	−7 to 31 days	52	69%	85%
Dobson and Carpenter (1991) NICU-treated	Acuity card	0.5	Binocular	−7 to 31 days	59	70%	90%
McDonald *et al.* (1986b) normals	Acuity card	1.0	Binocular	1–12 months	36	—	97%
Preston *et al.* (1987) visual abnormality	Acuity card	0.5	Binocular	2–8 months	20	95%	100%
McDonald *et al.* (1986a) normals	Acuity card	0.5	Binocular	18–36 months	36	92%	—
Heersema and van Hof-van Duin (1990) normals	Acuity card	0.3	Binocular	1–4 yr	50	82%†	92%
Hertz and Rosenberg (1988) cerebral palsy	Acuity card	0.5	Binocular	2–7 yr	59	69%	86%
Hertz (1988) mentally retarded	Acuity card	0.5	Binocular	8–17 yr	44	—	89%
Hertz and Rosenberg (1992) cerebral palsy	Acuity card	0.5	Binocular	1–8 yr	78	—	79%
Hertz *et al.* (1988) mentally retarded cortically visually impaired	Acuity card	0.5	Binocular	2–12 yr	8	25%	75%
Marx, Werner, Cohen-Mansfield and Hartmann (1990) non-communicative elderly	Acuity card	0.5	Binocular	74–96 yr	9	100%	100%
Mash *et al.* NICU-treated	Acuity card	0.5	Monocular	4–48 months	1918	67%	87%
McDonald *et al.* (1986b) normals	Acuity card	1.0	Monocular	1–12 months	66	—	86%
Dobson *et al.* (1990) NICU-treated	Acuity card	0.5	Monocular	4–12 months	382	58%	80%
Preston *et al.* (1987) visual abnormality	Acuity card	0.5	Monocular	2–8 months	40	88%	95%
Dobson and Carpenter (1991) NICU-treated	Acuity card	0.5	Monocular	4–24 months	1015	63%	85%
McDonald *et al.* (1986a) normals	Acuity card	0.5	Monocular	18–36 months	72	88%	—
Atkinson *et al.* (1982) normals	FPL	0.7	Monocular	4 months	12	—	100%
Maurer *et al.* (1989) normals	FPL	0.5	Monocular	6–12 months	57	86%	93%
Maurer *et al.* (1989) normals	FPL	0.5	Monocular	18–36 months	135	86%	96%
Maurer *et al.* (1989) aphakes	FPL	0.5	Monocular	6–36 months	101	57%	79%
Mash *et al.* NICU-treated	Acuity card	0.5	Interocular acuity difference	1 month	881	54%	76%
Preston *et al.* (1987) visual abnormality	Acuity card	0.5	Interocular acuity difference	2–8 months	20	75%	95%

*Number of interobserver test pairs.
†Percent of tests that differed by no more than 0.3 octave.

interobserver agreement was as high for these children as for children without severe abnormalities. This suggests that inaccurate TAC results are no more likely to occur in an infant or child with an ocular or neurological abnormality than in a normal infant or child. Our finding that interobserver agreement is as good in the eye tested second as in the eye tested first suggests that any fatigue associated with testing is unlikely to affect the reliability of acuity results. Age at testing had little affect on interobserver agreement, except that interobserver agreement was significantly lower at the difficult-to-test age of 24 months than at 48 months, when children tend to be more cooperative.

Prior to our analyses, we hypothesized that two test characteristics–long test duration and low tester confidence—would be indicators of inaccurate acuity results. However, there was little evidence that either long test duration or low observer confidence was predictive of low interobserver agreement.

One test characteristic that appeared to have an influence on acuity results was the subset of acuity cards used during testing. Acuity scores tended to be lower for subsets containing lower spatial frequency gratings and higher for subsets containing higher spatial frequency gratings. Anecdotally, some observers reported a tendency to be more satisfied with their decision to stop testing if they have presented several cards. This means that a tester using a high spatial frequency set of cards might tend to "push" the child to (or even past) threshold, whereas a tester using a low spatial frequency set of cards might tend to stop before threshold is reached, especially if the child is fussy or difficult to test.

This card subset effect may have contributed to some cases of poor interobserver agreement, as suggested by the finding that, at some test ages, interobserver agreement was influenced by the degree of disparity between subsets used in the test pairs. In clinical TAC testing, the tester is usually not masked to the spatial frequencies of the gratings used during testing, and card subsets are not used during testing. However, there may be a tendency for testers to overestimate acuity in eyes with acuity so poor that only a few cards can be presented prior to reaching threshold, and a tendency for testers to quit before threshold is reached in eyes with acuity that is better than what is expected based on clinical history.

Because previous studies have shown that some TAC (Quinn *et al.*, 1993) or FPL (Teller *et al.*, 1992) testers tend to be biased toward higher or lower acuity scores, we hypothesized that tester differences could have influenced interobserver agreement. However, a significant difference among testers in acuity scores was observed only at the 17-month test age, and there were no tester pairs that showed disproportionately high frequencies of poor agreement, as would have occurred if a tester with a substantial bias toward high acuity score had been paired with a tester with a bias toward lower acuity scores. Thus, tester differences were not a primary factor related to instances of poor interobserver agreement in the present study.

What can we conclude about why disagreement of more than 0.5 octave was found in 33% of test–retest comparisons and in 46% of IAD comparisons, and disagreement of more than one octave was found in 13% of test–retest comparisons and in 24% of IAD comparisons? Most likely, disagreement arises not from one single factor, but from a combination of factors, including age of the examinee, how many cards are presented before the child's acuity threshold is reached, and perhaps differences among testers in the criteria used for acuity estimation.

CONCLUSIONS

In conclusion, the results of the present study indicate that TACs provide acuity estimates that are as reliable as those obtained with previously-tested prototype acuity cards, as well as with other preferential-looking procedures. In the small proportion of cases in which an inaccurate acuity estimate is obtained, the present results suggest that the inaccurate estimate is more likely to be an underestimation of the true acuity value than an overestimation of that value. However, our analyses indicate that there is no reliable *single* indicator than an inaccurate estimate has been obtained. It may be that there are several factors that jointly predict inaccurate acuity results, but this could not be evaluated with our data set.

Thus, although the reliability of the TAC procedure in a clinical population is as high as that of other preferential looking measures of acuity, the results of a single acuity test should be interpreted in conjunction with other clinical findings from the patient, and when a question arises, a retest of acuity should be conducted.

REFERENCES

Atkinson, J., Braddick, O. & Pimm-Smith, E. (1982). 'Preferential looking' for monocular and binocular acuity testing of infants. *British Journal of Ophthalmology, 66*, 264–268.

Birch, E. E. (1985). Infant interocular acuity differences and binocular vision. *Vision Research, 25*, 571–576.

Birch, E. E. & Hale, L. A. (1988). Criteria for monocular acuity deficit in infancy and early childhood. *Investigative Ophthalmology and Visual Science, 29*, 636–643.

Dobson, V. & Carpenter, N. (1991). Interobserver agreement and observer bias in the acuity card procedure. In *Technical digest on noninvasive assessment of the visual system* (Vol. 1, pp. 20–23). Washington, D.C.: Optical Society of America.

Dobson, V. & Luna, B. (1993). Prototype and Teller Acuity Cards yield similar acuities in infants and young children despite stimulus differences. *Clinical Vision Sciences, 8*, 395–400.

Dobson, V., Carpenter, N. A., Bonvalot, K. & Bossler, J. (1990). The acuity card procedure: Interobserver agreement in infants with perinatal complications. *Clinical Vision Sciences, 6*, 39–48.

Hainline, L., Evelyn, L. & Abramov, I. (1989). Acuity cards—what do they measure? *Investigative Ophthalmology and Visual Science (Suppl.), 30*, 310.

Heersema, D. J. & van Hof-van Duin, J. (1990). Age norms for visual acuity in toddlers using the acuity card procedure. *Clinical Vision Sciences, 5*, 167–174.

Hertz, B. G. (1987). Acuity card testing of retarded children. *Behaviour and Brain Research, 24*, 85–92.

Hertz, B. G. (1988). Use of the acuity card method to test regarded

children in special schools. *Child Care and Health Development, 14*, 189–198.

Hertz, B. G. & Rosenberg, J. (1988). Acuity card testing of spastic children: Preliminary results. *Journal of Pediatrics and Ophthalmological Strabismus, 25*, 139–144.

Hertz, B. G. & Rosenberg, J. (1992). Effect of mental retardation and motor disability on testing with visual acuity cards. *Developmental Medicine and Child Neurology, 34*, 115–122.

Hertz, B. G., Rosenberg, J., Sjo, O. & Warburg, M. (1988). Acuity card testing of patients with cerebral visual impairment. *Developmental Medicine and Child Neurology, 30*, 632–637.

Marx, M. S., Werner, P., Cohen-Mansfield, J. & Hartmann, E. E. (1990). Visual acuity estimates in noncommunicative elderly persons. *Investigative Ophthalmology and Visual Science, 31*, 593–596.

Maurer, D., Lewis, T. L. & Brent, H. P. (1989). The effects of deprivation on human visual development: Studies of children treated for cataracts. In Morrison, F. J., Lord, C. E. & Keating, D. P. (Eds), *Applied developmental psychology* (Vol. 3, pp. 139–227). San Diego, Calif.: Academic Press.

McDonald, M. A., Ankrum, C., Preston, K., Sebris, S. L. & Dobson, V. (1986a). Monocular and binocular acuity estimation in 18- to 36-month-olds: Acuity card results. *American Journal of Optometry and Physiological Optics, 63*, 181–186.

McDonald, M. A., Sebris, S. L., Mohn, G., Teller, D. Y. & Dobson, V. (1986b). Monocular acuity in normal infants: The acuity card procedure. *American Journal of Optometry and Physiological Optics, 63*, 127–134.

McDonald, M. A., Dobson, V., Sebris, S. L., Baitch, L., Varner, D. & Teller, D. Y. (1985). The acuity card procedure: A rapid test of infant acuity. *Investigative Ophthalmology and Visual Science, 26*, 1158–1162.

Mohn, G., van Hof-van Duin, J., Fetter, W. P. F., de Groot, L. & Hage, M. (1988). Acuity assessment of non-verbal infants and children: Clinical experience with the acuity card procedure. *Developmental Medicine and Child Neurology, 30*, 232–244.

Preston, K. L., McDonald, M. A., Sebris, S. L., Dobson, V. & Teller, D. Y. (1987). Validation of the acuity card procedure for assessment of infants with ocular disorders. *Ophthalmology, 94*, 644–653.

Quinn, G. E., Berlin, J. A. & James, M. (1993). The Teller acuity card procedure—three testers in a clinical setting. *Ophthalmology, 100*, 488–494.

Robinson, J., Moseley, M. J. & Fielder, A. R. (1988). Grating acuity cards: Spurious resolution and the 'edge artifact'. *Clinical Vision Sciences, 3*, 285–288.

Sebris, S. L., Dobson, V., McDonald, M. A. & Teller, D. Y. (1987). Acuity cards for visual acuity assessment of infants and children in clinical settings. *Clinical Vision Sciences, 2*, 45–58.

Teller, D. Y., Mar, C. & Preston, K. L. (1992). Statistical properties of 500-trial infant psychometric functions. In Werner, L. A. & Rubel, E. W. (Eds), *Developmental psychoacoustics* (pp. 211–227). Washington, D. C.: American Psychological Association.

Teller, D. Y., McDonald, M. A., Preston, K., Sebris, S. L. & Dobson, V. (1986). Assessment of visual acuity in infants and children: The acuity card procedure. *Developmental Medicine and Child Neurology, 28*, 779–789.

Acknowledgements—The authors thank Robert D. Guthrie, M.D., Kathleen Godfrey, R.N., Lisa Getz, B.S., and the staff of the Magee-Womens Hospital Neonatal Intensive Care Unit for assistance in recruiting subjects, and Davida Teller, Ph.D. and Beatriz Luna, M.A. for comments on earlier versions of this paper. This research was supported by NIH grant EY 05804 (V.D.) and by the Magee-Womens Hospital Research Fund.